Towards Integrated Neural-Symbolic Systems for Human-Level AI: Two Research Programs Helping to Bridge the Gaps

Tarek R. Besold[a,*], Kai-Uwe Kühnberger[a]

[a]*Institute of Cognitive Science, University of Osnabrück, D-49069 Osnabrück, Germany*

Abstract

After a Human-Level AI-oriented overview of the status quo in neural-symbolic integration, two research programs aiming at overcoming long-standing challenges in the field are suggested to the community: The first program targets a better understanding of foundational differences and relationships on the level of computational complexity between symbolic and subsymbolic computation and representation, potentially providing explanations for the empirical differences between the paradigms in application scenarios and a foothold for subsequent attempts at overcoming these. The second program suggests a new approach and computational architecture for the cognitively-inspired anchoring of an agent's learning, knowledge formation, and higher reasoning abilities in real-world interactions through a closed neural-symbolic acting/sensing–processing–reasoning cycle, potentially providing new foundations for future agent architectures, multi-agent systems, robotics, and cognitive systems and facilitating a deeper understanding of the development and interaction in human-technological settings.

Keywords: research program, neural-symbolic integration, complexity theory, cognitive architectures, agent architectures

*Corresponding author
Email address: tarek.besold@uni-osnabrueck.de (Tarek R. Besold)

1. A Tale of Symbols and Signals: The Quest for Neural-Symbolic Integration

> "*I repeat my belief that learning has to be at the center of the artificial intelligence enterprise. While I do not regard intelligence as a unitary phenomenon, I do believe that the problem of reasoning from learned data is a central aspect of it.*" (Leslie Valiant, Valiant (2013), p. 163)

A seamless coupling between learning and reasoning is commonly taken as basis for intelligence in humans and, in close analogy, also for the biologically-inspired (re-)creation of human-level intelligence with computational means. Still, one of the unsolved methodological core issues in human-level AI, cognitive systems modelling, and cognitive and computational neuroscience—and as such one of the major obstacles towards solving the Biologically Inspired Cognitive Architectures (BICA) challenge (Samsonovich (2012))—is the question for the integration between connectionist subsymbolic (i.e., "neural-level") and logic-based symbolic (i.e., "cognitive-level") approaches to representation, computation, (mostly subsymbolic) learning, and (mostly symbolic) higher-level reasoning.

AI researchers working on the modelling or (re-)creation of human cognition and intelligence, and cognitive neuroscientists trying to understand the neural basis for human cognition, have for years been interested in the nature of brain-computation in general (see, e.g., Adolphs (2015)) and the relation between subsymbolic/neural and symbolic/cognitive modes of representation and computation in particular (see, e.g., Dinsmore (1992)). The brain has a neural structure which operates on the basis of low-level processing of perceptual signals, but cognition also exhibits the capability to efficiently perform abstract reasoning and symbol processing; in fact, processes of the latter type seem to form the conceptual cornerstones for thinking, decision-making, and other (also directly behavior-relevant) mental activities (see, e.g., Fodor & Pylyshyn (1988)).

Building on these observations—and taking into account that hybrid systems loosely combining symbolic and subsymbolic modules into one architecture turned out to be insufficient for the purpose—agreement on the need for fully integrated neural-cognitive processing has emerged (see, e.g., Bader & Hitzler (2005); d'Avila Garcez et al. (2015)). This has several reasons also beyond the analogy to the described functioning principles of the brain:

- In general, network-based approaches possess a higher degree of biological motivation than symbol-based approaches, also outmatching the latter in terms of learning capacities, robust fault-tolerant processing, and generalization to similar input. Also, in AI applications they often enable flexible tools (e.g., for discovering and processing the internal structure of possibly large data sets) and efficient signal-processing models (which are biologically plausible and optimally suited for a wide range of applications).

- Symbolic representations are generally superior in terms of their interpretability, the possibilities of direct control and coding, and the extraction of knowledge when compared to their (in many ways still black box-like) connectionist counterparts.[1]

- From a cognitive modelling point of view, subsymbolic representations for tasks requiring symbolic high-level reasoning might help solving, among many others, the problem with "too large" logical (epistemic) models (see, e.g., Gierasimczuk & Szymanik (2011)) which seem to lead to implausible computations from the reasoning agent's perspective (Degremont et al.

[1] Based on results as, for instance, the ones presented in Olden & Jackson (2002), it has been argued that the inner mechanics of artificial neural networks (ANNs) can be made accessible using randomization methods and similar. While this is true when seeing ANNs as quantitative tools or means of statistical modelling, from the quite different perspective of mechanistic or explanatory knowledge about principles, rules, and processes within ANNs as part of cognitive architectures the black box character remains (with rule extraction methods, as, e.g., proposed in Andrews et al. (1995), d'Avila Garcez et al. (2001), or Zhou et al. (2003), mitigating the problem only to a minimal degree).

(2014)). On the other hand, being able to lift subsymbolic brain-inspired models and corresponding simulations to a symbolic level of description and analysis promises to close the interpretative and explanatory gap between actual biologically-motivated model dynamics and observed behavior also for tasks involving complex or abstract reasoning.

In summary, cognitive-level interpretations of artificial neural network (ANN) architectures and accurate and feasible neural-level models of symbolic processing are highly desirable: as an important step towards the computational (re-)creation of mental capacities, as possible sources of an additional (bridging) level of explanation of cognitive phenomena of the human brain (assuming that suitably chosen ANN models correspond in a meaningful way to their biological counterparts), and also as important part of future technological developments (also see Sect. 6).

But while there is theoretical evidence indicating that both paradigms indeed share deep connections, how to explicitly establish and exploit these correspondences currently remains a mostly unsolved question. In the following, after a concise overview of the state of the art in the field of neural-symbolic integration in Sect. 2, as an invitation to researchers from the relevant communities two research programs are laid out which have the potential to shed light on this foundational issue: The first one, summarized in Sect. 3, targets a better understanding of the empirical differences and commonalities between formalisms from the symbolic and the subsymbolic paradigm on the level of computational complexity in more scenario-specific and fine-grained ways than previously achieved. The second one, outlined in Sect. 4, gives a conceptual sketch of a research effort developing a new approach and computational architecture for the cognitively-inspired anchoring of an agent's learning, knowledge formation, and higher reasoning abilities in real-world interactions through a closed neural-symbolic acting/sensing–processing–reasoning cycle. If implemented successfully, the second program will lay the foundations for a new generation of intelligent agent systems, also giving evidence of the capacities of

fully integrated neural-symbolic learning and reasoning on system level. Thus, as explained in Sect. 5, when taken together both programs—besides significantly advancing the field of neural-symbolic integration—promise to greatly contribute to all four pillars and the respectively associated main scientific views of BICA identified in Stocco et al. (2010). Additionally, major impact of the research programs (and the corresponding form of neural-symbolic integration) can also be expected on an immediate technological level in the area of smart systems. Sect. 6 sketches the corresponding technological scenario and describes an envisioned example from the domain of ambient-assisted living (AAL).

2. Status Quo in Neural-Symbolic Integration as of 2015

Concerning our current understanding of the relationship and differences between symbolic and subsymbolic computation and representation, the cognitive-level "symbolic paradigm" is commonly taken to correspond to a Von Neumann architecture (with predominantly discrete and serial computation and localized representations) and the neural-level "subsymbolic paradigm" mainly is conceptualized as a dynamical systems-type approach (with distributed representations and predominantly parallel and continuous computations).

This divergence notwithstanding, both symbolic/cognitive and subsymbolic/neural models in theory are considered substantially equivalent in most (if not all) practically relevant dimensions (see Sect. 2.1 for details). Still, in general experiences from application studies consistently and reliably show different degrees of suitability and performance of the paradigms in different types of application scenarios, with subsymbolic approaches offering themselves, e.g., for effective and efficient solutions to tasks involving learning and generalization, while high-level reasoning and concept composition are commonly addressed in symbolic frameworks. Unfortunately, general explanations (and solutions) for this foundational dichotomy this far have been elusive when using standard methods of investigation.

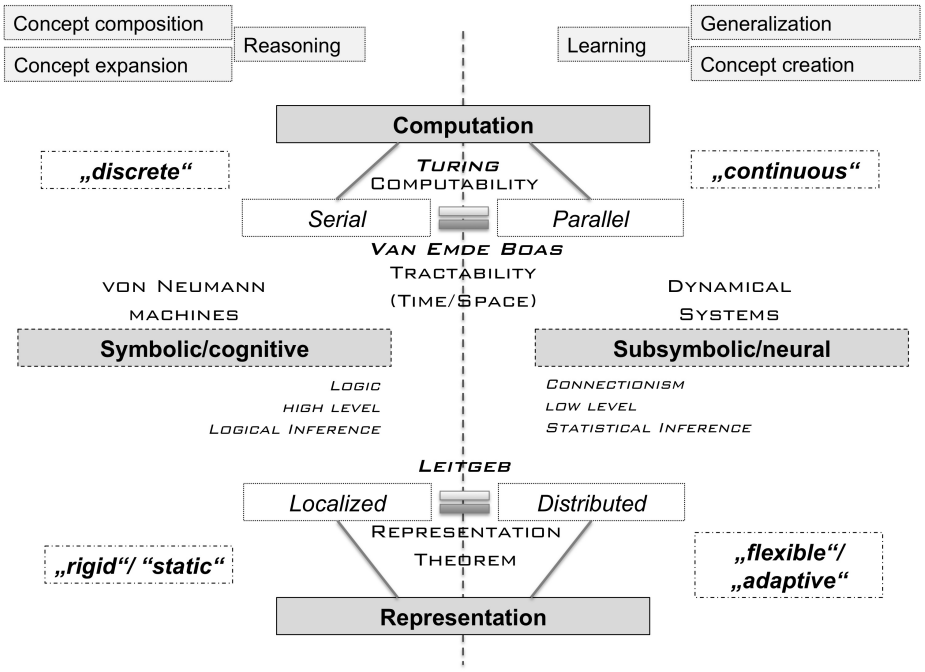

Figure 1: A schematic overview of common conceptualizations concerning symbolic and subsymbolic representation and computation, as well as the connections and differences between both paradigms (see Sect. 2.1 for details concerning the indicated formal equivalences).

2.1. Formal Analysis of Symbolic and Subsymbolic Computation and Representation

According to our current knowledge, from a formal perspective—especially when focusing on actually physically-realizable and implementable systems (i.e., physical finite state machines) instead of strictly abstract models of computation, together with the resulting physical and conceptual limitations—both symbolic/cognitive and subsymbolic/neural models seem basically equivalent.

Notwithstanding partially differing theoretical findings and discussions (as, e.g., given in Tabor (2009)), both paradigms are considered computability-equivalent in practice (Siegelmann (1999)). Also from a tractability perspective, for instance in van Rooij (2008), equivalence in practice with respect to classical dimensions of analysis (i.e., interchangeability except for a polynomial overhead) has been established, complementing and supporting the theoretical suggestion of equivalence by Van Emde Boas in his Invariance Thesis (van Emde Boas (1990)) . Finally, Leitgeb provided an *in principle* existence result in Leitgeb (2005), showing that there is no substantial difference in representational or problem-solving power between dynamical systems with distributed representations or symbolic systems with non-monotonic reasoning capabilities.

Still, these results are only partially satisfactory: Although introducing basic connections and mutual dependencies between both paradigms, the respective levels of analysis are quite coarse and the found results are only existential in character. While establishing the *in principle* equivalence described above, in Leitgeb (2005) no constructive methods for how to actually obtain the corresponding symbolic counterpart to a subsymbolic model (and vice versa) are given.

Concerning the complexity and computability equivalences, while the latter is supported by the results in Leitgeb (2005), the former stays mostly untouched: While coming to the same conclusion, i.e., the absence of *substantial* differences between paradigms (i.e., differences at the level of tractability classes), no further clarification or characterization of the precise nature and properties of the polynomial overhead between symbolic and subsymbolic approaches is provided.

2.2. Neural-Symbolic Integration in AI, Cognitive Modelling, and Machine Learning

Research on integrated neural-symbolic systems (especially in AI and to a certain extent also in cognitive modelling) has made significant progress over the last two decades (see, e.g., Bader & Hitzler (2005); d'Avila Garcez et al. (2015)); partially, but not exclusively, in the wake of the development of deep learning approaches to machine learning (see, e.g. Bengio et al. (2013); Schmidhuber (2015)). Generally, what seem to be several important steps towards the development of integrated neural-symbolic models have been made:

- From the symbolic perspective on the capacities of subsymbolic computation and representation, the "Propositional Fixation" (i.e., the limitation of neural models on implementing propositional logic at best) has been overcome, among others, in models implementing modal or temporal logics with ANNs (see, e.g., d'Avila Garcez et al. (2008)).

- From the subsymbolic perspective, neural computation has been equipped with features previously (almost) exclusively limited to symbolic models by adding top-down governing mechanisms to modular, neural learning architectures, for example, through the use of "Conceptors" (Jaeger (2014)) as computational principle.

- Deep learning approaches to machine learning—by the high number of parameterized transformations performed in the corresponding hierarchically structured models—seem to, at first sight, also conceptually provide what can be interpreted as different levels of abstraction above and beyond mere low-level processing. The resulting networks partially perform tasks classically involving complex symbolic reasoning such as, for instance, the labeling of picture elements or scene description (see, e.g., Karpathy & Fei-Fei (2014); Vinyals et al. (2014)).

- Recently proposed classes of subsymbolic models such as "Neural Turing Machines" (Graves et al. (2014)) or "Memory Networks" (Weston et al.

(2015)) seem to also architecturally narrow the gap between the (subsymbolic) dynamical systems characterization and the (symbolic) Von Neumann architecture understanding.

Nonetheless, all these developments (including deep neural networks as layered recurrent ANNs) stay within the possibilities and limitations of the respective classical paradigms without significantly changing the basic formal characteristics of the latter.

2.3. Summary

Although remarkable successes have been achieved within the respective paradigms, the divide between the paradigms persists, interconnecting results still either only address specific and non-generalizable cases or are *in principle* and non-constructive, benchmark scenarios for principled comparisons (e.g., in terms of expressive strength of knowledge representation formalisms or descriptive complexity) between subsymbolic and symbolic models have still not been established, and questions concerning the precise nature of the relationship and foundational differences between symbolic/cognitive and subsymbolic/neural approaches to computation and representation still remain unanswered (see, e.g., Isaac et al. (2014)): in some cases due to a lack of knowledge for deciding the problem, in others due to a lack of tools and methods for properly specifying and addressing the relevant questions.

3. Identifying and Exploring Differences in Complexity

Focusing on the just described lack of tools and methods, together with the insufficient theoretical knowledge about many aspects of the respective form(s) of computation and representation, in the first of the two envisioned research programs (initially introduced in Besold (2015)), the classical findings concerning the relation and integration between the symbolic/cognitive and the subsymbolic/neural paradigm described in Sect. 2 shall be revisited in light of new

developments in the modelling and analysis of connectionist systems in general (and ANNs in particular), and of new formal methods for investigating the properties of general forms of representation and computation on a symbolic level.

To this end, taking into account the apparent empirical differences between the paradigms and (especially when dealing with physically-realizable systems) assuming basic equivalence on the level of computability, emphasis shall be put on identifying and/or developing adequate formal tools and investigating previously unconsidered aspects of existing equivalence results. Focus shall be put on the precise nature of the polynomial overhead as computational-complexity difference between paradigms: Most complexity results for symbolic/cognitive and subsymbolic/neural computations have been established using exclusively TIME and SPACE as classical resources (see, e.g., Thomas & Vollmer (2010); Sima & Orponen (2003)), and the tractability equivalence between paradigms (see, e.g., van Rooij (2008)) mostly leaves out more precise investigations of the remaining polynomial overhead. Against this background, the working hypotheses for the program are that TIME and SPACE are not always adequate and sufficient as resources of analysis for elucidating all relevant properties of the respective paradigms, and that there are significant characteristics and explanations to be found on a more fine-grained level than accessible by classical methods of analysis (settling on the general tractability level).

The main line of research can be summarized in two consecutive questions (corresponding to the stated working hypotheses), one starting out from a more subsymbolic, the other from a more symbolic perspective:

- **Question 1:** Especially when considering subsymbolic/neural forms of computation and the associated dynamical systems conception, the adequacy and exhaustiveness of the classical approaches to complexity analysis using only TIME and SPACE as resources for a fully informative characterization must be questioned. Are there more adequate resources which should be taken into account for analysis?

- **Question 2:** Especially when considering the symbolic level, are there more adequate approaches/methods of analysis available than classical complexity theory, allowing to take into account formalism- or calculus-specific characterizations of computations or to perform analyses at a more fine-grained level than tractability?

Finally, in an integrative concluding step taking into account the methods and findings resulting from the previous two, a third question shall be investigated:

- **Question 3:** Can the *in principle* results from Leitgeb (2005) be extended to more specific and/or constructive correspondences between individual notions and/or characterizations within the respective paradigms?

Answers to these questions (and the resulting refined tools and methods) promise to contribute to resolving some of the basic theoretical and practical tensions described in Sect. 1 and 2: Although both paradigms are theoretically undistinguishable (i.e., equivalent up to a polynomial overhead) in their general computational-complexity behavior using classical methods of analysis and characterization results, empirical studies and application cases using state of the art approaches still show clear distinctions in suitability and feasibility of the respective paradigms for different types of tasks and domains without us having an explanation for this behavior. Parts of this divergence might be explained by previously unconsidered and unaccessible complexity-related properties of the respective approaches and their connections to each other.

The targeted level of work is situated between the (purely theoretical) development of methods in complexity theory, network analysis, etc. and the (purely applied) study of properties of computational and representational paradigms by applying existing tools: Previous work from the different fields and lines of research shall be assessed and combined—in doing so, where necessary, adapting or expanding the respective methods and tools—into new means of analysis, which then shall subsequently be applied to suitably selected candidate models representing paradigmatic examples of symbolic or subsymbolic rep-

resentations/computations with respect to features relevant for the respective question(s) at hand.

3.1. Proposed Program Structure and Approaches

The envisioned research program is divided into three stages, corresponding to the three initially posed questions. Each of the latter can (and should) be addressed in its own right, but when taken together the respective answers promise to also shed light on the bigger question for the existence and the precise nature of foundational differences between the symbolic and the subsymbolic paradigm.

3.1.1. Adequate Resources for Analysis.

TIME and SPACE are the standard resources considered in classical complexity analyses of computational frameworks. Correspondingly, most results concerning complexity comparisons between symbolic and subsymbolic models of computation also focus on these two dimensions (as do, e.g., the aforementioned results in van Rooij (2008); van Emde Boas (1990)).

Still, the reading of TIME and SPACE as mostly relevant resources for complexity analysis is closely connected to a Turing-style conception of computation and a Von Neumann-inspired architecture as machine model, working, e.g., with limited memory. Especially when considering other computational paradigms with different characteristics, as, e.g., the dynamical systems model commonly associated to the subsymbolic/neural paradigm, the exhaustiveness and adequateness of TIME and SPACE for a full analysis of all relevant computational properties has to be questioned. Instead, it seems likely that additional resources specific to the respective model of computation and architecture have to be taken into account in order to provide a complete characterization.

Thus, in a first stage of the program, popular network types on the subsymbolic/neural side shall be investigated for relevant dimensions of analysis other than TIME and SPACE. Besides the classical standard and recurrent approaches, of course also other models such as recurrent spiking neural networks

(see, e.g. Gerstner et al. (2014)), Long Short-Term Memory networks and extensions thereof (see, e.g., Monner & Reggia (2012)), or recurrent stochastic neural networks in form of Boltzmann machines (Ackley et al. (1985)) and restricted Boltzmann machines (Hinton (2002)) could (and eventually will have to) be considered.

Taking recurrent networks of spiking neurons as examples, also measures such as spike complexity (a bound for the total number of spikes during computation; Uchizawa et al. (2006)), convergence speed (from some initial network state to the stationary distribution; Habenschuss et al. (2013)), sample complexity (the number of samples from the stationary distribution needed for a satisfactory computational output; Vul et al. (2014)), or network size and connectivity seem to be promising candidates for relevant dimensions of analysis.

These and similar proposals for the other network models shall be critically assessed and, where possible, put into a correspondence relation with each other, allowing to meaningfully generalize between different subsymbolic/neural models and to provide general characterizations of the respective computations. Having in mind the overall goal of connecting subsymbolic and symbolic approaches, a guiding heuristic for the selection of candidate proposals and also during the final step of cross-model generalization is provided by the degree of expected cross-paradigmatic relevance: Taking examples from above, while spike complexity—due to its direct correspondence to energy consumption during computation—by itself seems to be an interesting and (especially biologically) highly relevant perspective for characterizing the resource-requirements of computations in a recurrent spiking neural network, its relevance for characterizing the complexity of this type of ANN as compared to a corresponding symbolic model might be limited due to the lack of a direct counterpart to the concept of energy use in the logic-based framework. Convergence speed, on the other hand, while (although less directly) still related as resource to energy consumption in the network setting, might allow for a more direct and adequate bridging to symbolic forms of computation, possibly corresponding to concepts such as the required number of inference steps in the calculus of a

certain logic-based formalism.

At the end of this stage, new proposals for adequate resources usable in refined complexity analyses for subsymbolic/neural computation, together with application examples in terms of proof of concept analyses of popular paradigms, will be available.

3.1.2. Adequate Methods of Analysis

In parallel to and/or following the search for more adequate resources for complexity analyses of mostly subsymbolic/neural models of computation, in a second stage of the program emphasis shall be shifted towards the symbolic/cognitive side. While staying closer to the classical conception of complexity in terms of TIME and SPACE, recent developments in different fields of theoretical computer science shall be combined into tools for more model-specific and fine-grained analyses of computational properties.

Parameterized-complexity theory (see, e.g., Downey & Fellows (1999)) makes the investigation of problem-specific complexity characteristics possible, while tools such as, e.g., developed in the theory of proof-complexity (see, e.g., Krajícek (2005)) allow for more varied formalism- or calculus-specific characterizations of the respective computations than currently done. Additionally, tools from descriptive complexity theory (see, e.g., Immerman (1999)) and work from model-theoretic syntax (see, e.g., Rabin (1965)) seem likely to offer chances for shedding light on complexity distinctions below the tractability threshold (i.e., for exploring the precise nature of the polynomial overhead) and to allow for more fine-grained and discriminative comparisons between paradigms and models.

Thus, results from the just mentioned fields/techniques can be examined for their applicability to better characterizing symbolic computation and to potentially establishing conceptual connections to characterizations of subsymbolic computation from the previous stage. Taking into account their specific properties and strengths, the corresponding tasks for the respective approaches can be summarized as follows:

- Parameterized-complexity theory: Taking into account problem- and application-

specific properties of (families of) problems and connecting these to results describing specific properties of subsymbolic or symbolic computation and representation, try to explain the different suitability of one or the other paradigm for certain types of tasks.

- Descriptive complexity theory and model-theoretic syntax: Attempt to explore complexity distinctions between different forms of symbolic and between symbolic and subsymbolic computation also in more fine-grained ways than by mere tractability considerations (e.g., also taking into account the polynomial-time hierarchy and the logarithmic-time hierarchy).

- Proof-complexity theory: Explore formalism- and calculus-specific properties of symbolic computations and try to map these onto properties of specific subsymbolic models.

Concerning a concrete implementation, one possibility is to initially apply methods and ideas from parameterized-complexity theory to existing and accepted computational complexity results concerning subsymbolic and symbolic computation and representation in certain tasks and domains. Here, task- and domain-specific properties of the respective paradigms shall be investigated also beyond and below the level of classical tractability, attempting to elucidate parts of the reasons for the empirical differences between approaches.

In a first step, previous results from the literature can be taken and the parameterized dimension of analysis can be added, so that specificities of the task or domain can be investigated while still maintaining the connection to previous findings and the embedding in a more general scientific context. Subsequently, the overall approach of parameterized analysis can be combined with notions taken from or inspired by the other aforementioned forms of (originally symbolic-focused) fine-grained complexity analysis which allow to discriminate complexity properties below the classical tractability threshold and on a formalism- and calculus-specific level. The resulting combined approaches can then be applied to specific symbolic models selected based on hypotheses and correspondences

obtained in the previous stage of the program (described in Sect. 3.1.1), as well as on their suitability for the type of analysis under consideration.

At the end of this stage, proposals for refined methods of analysis especially for forms of symbolic/cognitive computation and application examples in terms of proof of concept analyses, together with suggestions for correspondences to models of subsymbolic/neural computation, will be available.

3.1.3. Correspondences Between Paradigms

In a third and final part of the program, by combining the results of the preceding stages, additional dimensions can be added to previous analyses and established equivalence results, and the precise nature of the polynomial overhead as computational difference between paradigms can better be explained. Also, the outcomes of previous stages shall be integrated where meaningfully possible, ideally providing the foundations for a general set of refined means of analysis for future comparative investigations of symbolic/cognitive and subsymbolic/neural computation.

Depending on previous outcomes, some of the following (interrelated) questions are expected to be addressable:

- Given the *in principle* equivalence between (symbolic) non-monotonic logical systems and (subsymbolic) dynamical systems, is it possible to establish complexity-based systematic conceptual relationships between particular logical calculi and different types of subsymbolic networks?

  If such relationships can indeed be identified, this will be informative in at leas two ways: On a functional level, given that certain subsymbolic or symbolic approaches are known to perform well on specific tasks, knowledge about correspondences between paradigms can narrow down the range of candidates for solving the same tasks using methods from the respective other paradigm. On a structural level, the envisioned conceptual relationships promise additional ways of comparing the respective structure of the conceptual spaces of subsymbolic and symbolic approaches—

while their internal organization (e.g., concerning gradual differences in expressivity or computational properties between different logics, or learning capacity or computational complexity between different types of ANNs) by now has been mapped out fairly well, establishing correspondences and transferring known structural orderings from the respectively better-known space to the other still poses major challenges.

- Can adaptations in network structure and/or (the artificial equivalent of) synaptic dynamics (see, e.g., Choquet & Triller (2013)) in a neural representation in a systematic way be related to re-representation in a logic-based representation, or (alternatively) is there a systematic correspondence on the level of change of calculus? Can changes in network type in a neural representation in a systematic way be related to changes of non-monotonic logic in a symbolic representation?

  As dynamic adaptations of network topology, connection properties and/or synaptic properties can be taken as hallmarks of the functioning of the human brain and, to a large extent, also of many successful ANN models, finding or creating corresponding mechanisms for symbolic approaches would promise to also allow for a transfer of functional properties from the subsymbolic to the symbolic paradigm, for instance, with regard to applications in learning and generalization. On the other hand, by observing the corresponding changes on the symbolic level a better understanding and explanation of the actual functioning and mechanisms at work in the subsymbolic case can be expected. Similar expectations can be maintained on the less biologically-motivated, but from a computer science perspective currently possibly even more relevant level of network types and different logics.

- Can the correspondences and differences between novel network models approximating classical symbolic capacities (as, e.g., top-down control) or architectures (as, e.g., a Von Neumann machine) and the original symbolic concepts be characterized in a systematic way?

While several promising network models offering partial interpretations in symbolic terms have recently been proposed (see Sect. 2.2), this far the correspondences between the new model and the classical notions have mostly been established on a case-by-case basis and no targeted development aiming at methodically developing the newly introduced approaches further towards fully covering the classical conceptualizations have been presented. Here, systematic correspondences could offer guidance for the corresponding process.

At the end of this stage, partial answers to some of the stated questions together with proposals for future lines of investigation continuing the work started in the program will be available. Also, suggestions for new tools and methods for the comparative analysis of symbolic/cognitive and subsymbolic/neural computation will be made.

3.2. Expected Outcomes

If implemented successfully, the sketched research program is expected to be highly beneficial for neural-symbolic integration on at least two dimensions, a methodological and a theoretical one.

From the methodological point of view, new general approaches and updated and refined formal tools for better and more adequately analyzing and characterizing the nature and mechanisms of representation and computation in the corresponding paradigm(s) will be developed: Alternative resources complementing TIME and SPACE for the characterization of properties of (especially subsymbolic/neural) computation will be provided, and emphasis will be put on making model-specific properties of the respective computing mechanisms accessible. Also, alternative methods complementing the classical complexity-theoretical approach to the characterization of properties of (especially symbolic/cognitive) computation will be explored and canonized. Here, the focus will be on opening up formalism- or calculus-specific properties of the respective computing mechanisms, and on offering more fine-grained insights than available in the classical framework.

From the theoretical point of view, new perspectives on the relation between symbolic/cognitive and subsymbolic/neural representation and computation will be explored and a better understanding of the respective approach(es) and their interaction (with a strong orientation towards a future integration of conceptual paradigms, of levels of explanation, and of involved scientific disciplines) shall be established. Emphasis will be put on understanding the interaction between model-specific changes in one paradigm and corresponding adaptations of the respective conceptual or formal counterpart within the other paradigm.

4. Anchoring Knowledge in Interaction in a Framework and Architecture of Computational Cognition

The research program proposed in the previous section aims at uncovering basic distinctions and connections between subsymbolic and symbolic computation and representation on a—although strongly empirically motivated—mostly theoretical level. Complementing and completing this approach, in the following (building on parts of a larger proposal originally presented in Besold et al. (2015)) a second research endeavor is outlined, aiming at integrating neural-level and cognitive-level approaches in a new perspective and cognitive system architecture for interaction-grounded knowledge acquisition and processing in a closed acting/sensing–processing–reasoning cycle. In addition to the question of neural-symbolic integration it, thus, also is of direct relevance for practical challenges such as representational re-description and the progressive acquisition of abstract representations from raw sensory inputs in robot architectures (Guerin et al. (2013)).

4.1. An Agent's Knowledge for/in/from Its World

Natural agents in many situations in their reasoning seem to rely on an enormous richness of representations (multimodal, grounded, embodied and situated), with many layers of representation at different levels of abstraction, together with dynamic re-organization of knowledge. Also, real-world situations

require agents to perform what can be interpreted as dynamic changes or alignments of representation, as different agents might use different languages and levels of description. Unfortunately, when trying to follow the natural example by transferring and (re-)creating this representational richness and diversity in artificial agents, the resulting mismatches cannot be cured by standardization, but arise due to differences in the environment, tasks to be solved, levels of abstraction, etc. Additionally, real-world applications also demand online and bidirectional learning that takes place in real-time, as well as the adaptation to changes in the environment, to the presence of new agents, and to task changes.

In order to be able to face these challenges, we envision a system operating on different levels of representations (corresponding to different formal layers in the system's architecture). The hierarchy could consist, for instance, of a (lowest) neural layer learning on the perception/motor level, an anchoring layer learning elementary (semi-)symbolic representations of objects, a reactive layer taking over in critical situations, a deep learning layer learning on more abstract levels, a symbolic layer doing reasoning and planning, and a (higher) symbolic layer providing the core ontology. Some of these layers have obvious, some have partial, some have fuzzy, and some have no mappings/relations between themselves.

Now, a corresponding architecture should be in a "pre-established harmony" with initial correspondences between and across levels: Triggering an abstract plan to move from A to B should result in the motor action to move from A to B, classifying on the neural level a certain perceptual input such as, for instance, a chair should result in the activation of the concept "chair" in the ontology or the working memory, and so on. And whilst the basic links might be hard coded, learning a new concept on the subsymbolic level should somehow result in a new concept entry in the ontology, i.e., there should be interaction between the different layers in terms of information and conceptualizations. Finally, when thinking about a simulated or actual system that is operating on these interacting levels in a multi-representational manner it should allow for the learning or detection of obvious mappings between the layers, for detecting novelties and

correlations, for systematically unfolding the specific properties of structures on different levels, or for finding invariant properties of the interactions between levels.

4.2. From Interaction to Knowledge and Back

Against this background, in Besold et al. (2015) a program has been proposed for 'anchoring knowledge in interaction', aiming at developing, theoretically and practically, a conceptual framework and corresponding architecture that model an agent's knowledge, thinking, and acting truly as interrelated parts of a unified cognitive capacity. That is, knowledge is seen as multi-layered phenomenon that appears at different levels of abstraction, promotes interaction between these levels of abstraction, is influenced by the interaction between agent and environment (potentially including other agents), and is essentially linked to actions, perception, thinking, and being. Thus, the future architecture aims to anchor and embody knowledge by the interaction between the agent and its environment (possibly including other agents), to give an approach to lift the resulting situated action patterns to a symbolic level, to reason by analogy on the abstract and the subsymbolic level, to adapt, or in case of clashes, repair the initial representations in order to fit to new situations, and to evaluate the approach in concrete settings providing feedback to the system in a reactive-adaptive evolutionary cycle. Among others, this will require a new paradigm for neural-symbolic knowledge repositories featuring different integrated levels and forms of knowledge representation (as, e.g., multi-modal or hybrid).

On an embodied level, elementary forms of representations shall be learned from an agent's interactions within an environment. As the resulting multi-modal representations are likely to be noisy, uncertain, vague, unstable over time, and represented in different languages in different agents, an extension of the anchoring framework in robotics Coradeschi & Saffiotti (2000) to grounding not only objects, but also certain general observable properties appearing in the environment, will be needed.

Once an interaction-based neural representation of knowledge has been ob-

tained, neural systems can promote robust learning from data, as part of an online learning and reasoning cycle. On this level, a lifting procedure shall be specified that will produce descriptions, thus lifting grounded situations and an agent's action patterns to a more abstract (symbolic) representation. This can be done using techniques from machine learning such as, e.g., deep neural networks (as mentioned in Sect. 2.2) and analogy-making across networks (i.e., representation systems) and learning processes (i.e., procedural approaches) as proposed in d'Avila Garcez et al. (2015).

Although one could consider the neural-symbolic part already as solved with the "syntactic" lifting of neural representations to symbol-based ones, the envisioned research program targets an additional "semantic" step: As already mentioned, initial multi-modal representations lifted from the subsymbolic level can be error-prone and are likely to be represented in different and possibly at first incompatible representation languages between different agents. In order to also close these contentual gaps within and between agents, a dynamic re-organization and alignment (based on ontology repair mechanisms, analogy, concept invention, and knowledge transfer) is foreseen. These mechanisms foster adaptation of an agent to new situations, the alignment between representations of different agents, the reformulation of knowledge entries, and the generation of new knowledge.

In summary, the envisioned account of the emergence of representations through cognitive principles in an agent (or multi-agent) setting can be conceptualized as follows: Grounding knowledge in cognitively plausible multimodal interaction paradigms; lifting grounded situations into more abstract representations; reasoning by analogy and concept blending at more abstract levels; repair and re-organization of initial and generated abstract representations.

4.3. Proposed Program Structure and Approaches

The proposed approach requires the integration of expressive symbolic knowledge representation formalisms, relational knowledge, variables, and first-order logic on the one hand with representations of sensorimotor experiences, action

patterns, connectionist representations, and multi-modal representations on the other—basically exhausting the entire spectrum of levels of representation considered in neural-symbolic integration.

With respect to the formalization, research methods from machine learning (e.g. cross-validation as described by Dietterich (1998) or Arnold et al. (2010)'s layer-wise model selection in deep neural networks) will be applied to learn conceptual knowledge from subsymbolic data. The resulting conceptual knowledge will be provided as input to the analogy-making process to generate new concepts by abstraction and transfer of knowledge in a domain-independent and multi-modal setting. As this might potentially change the signatures of the underlying representation language(s), the theory of institutions (Diaconescu (2008)) will be used in order to model the corresponding dynamic changes of languages. Finally, the repair of theories and the concept invention mechanisms will be linked to analogy-making and are situated on the level of higher-order logic (Bundy (2013); Lehmann et al. (2013)).

From the perspective of neural-symbolic integration, the envisioned research program can be subdivided into three main modules:

- Cognitive Foundations of Knowledge and Anchoring: Approaches from computational neuroscience and network-level cognitive modeling (as, e.g., the recently proposed framework of conceptors in dynamical system models; Jaeger (2014)), together with theoretical considerations on sensorimotor interactions as part of knowledge formation (Fischer (2012)), serve as basis for the creation of low-level input representations and content for the subsequent stages of processing and reasoning. These inputs are then used in an anchoring step (Coradeschi & Saffiotti (2000)) grounding symbols referring to perceived physical objects in the agent's environment. Compared to previous approaches (Chella et al. (2003)), in the present context anchoring shall be developed further and conducted under even more general conditions: If the proposed program is implemented successfully, among others, anchoring will happen top-down and bottom-up

during learning, and new symbols for new objects and categories are dynamically introduced by repair and concept invention mechanisms.

Although at first sight lying outside the core domain of neural-symbolic integration, an effective solution to the anchoring problem in the just described form seems nonetheless indispensable. A fully integrated neural-symbolic cognitive system should bridge from direct sensory inputs to high-level symbolic representations and vice versa, with the equivalence between representations not only residing on a syntactic but also on a semantic and computational level—the latter of which require the capacity to reason about perceptual input also on lower representation levels in conceptual terms.

- Lifting Knowledge from the Subsymbolic to the Symbolic Level: Deep learning as a form of representation learning that aims at discovering multiple levels of representation has shown promising results when applied to real-time processing of multimodal data (De Penning et al. (2011)), and state-of-the-art deep learning methods and algorithms have been able to train deep networks effectively when applied to different kinds of networks, knowledge fusion, and transfer learning (Bengio (2009)). However, more expressive descriptions and forms of representation have become more difficult to obtain from neural networks.

  This module constitutes the centerpiece from the perspective of neural-symbolic integration. In it, neural learning shall be combined with temporal knowledge representation in stochastic networks, for instance by using variations of the Restricted Boltzmann Machine model (Hinton (2012)). The resulting approach then will offer a method for validating hypotheses through the symbolic description of the trained networks whilst robustly dealing with uncertainty and errors through a Bayesian inference model. Furthermore, using the "conceptual spaces" from Gärdenfors (2000) (building and expanding upon work presented, e.g., in LeBlanc & Saffiotti (2008)), symbolic and subsymbolic data shall be linked in the

proposed complex loop of sensing, processing, and reasoning.

- Analogy/Blending and Concept Formation/Reformation: While the first module on cognitive foundations of knowledge and anchoring shall provide the basis on which the lifting process can operate, the envisioned framework is completed by a third module framing and supporting the lifting process on an upper representational level. In this part of the program, analogy-making and concept (re)formation shall be added to the acting/sensing–processing–reasoning cycle in order to model high-level knowledge processing and to provide feedback and partial guidance to the knowledge acquisition and interpretation processes on lower levels.

  Analogy is classically understood as a method to detect and operate on structural commonalities between two domains (Gentner et al. (2001)), and due to its central role in human cognition over the years a significant number of computational models of analogy-making have been developed in AI (Besold (2011)). The targeted approach in the sketched program advances beyond the current state of the art in that generalizability, multi-modal representations, and the grounding in the agent's interaction with the environment are considered to be essential features. Additionally, analogy-making shall not only happen on the (symbolic) knowledge level, but already before that during learning and knowledge lifting, leading to cross-informing learning processes between similar sensory settings. Furthermore, analogies shall directly be linked to knowledge repair and knowledge formation mechanisms in order to facilitate the resolution of errors appearing almost unavoidably as part of the described paradigm: An important way in which new concepts are formed is through the (analogy-based) combination of existing concepts into a new concept by a blending mechanism (Fauconnier & Turner (2002)), or by the evolution of existing concepts that have proved inadequate. Inadequacies of the latter type are often revealed by failures of inference using the old concepts. Here, based on the reformation algorithm (Bundy (2013)), generic

mechanisms for repairing agents' faulty representations (especially those produced by imperfect analogies) will be developed.

When taken together, solving all three modules allows for a completely integrated neural-symbolic architecture, bridging not only on a syntactic level from connectionist to symbolic representations, but also taking into account semantic structures on all levels, from regularities and governing rules in the perceived environment of an agent (accessed via computations on the sensory input stream), through commonalities on an intermediate procedural level, to similarity structures and concept (re)formation on abstract knowledge entries. In doing so, the sketched cognitive computational framework will come closer to its biological inspiration, the human brain and mind, in functional and (abstracted) structural terms than previous architectures, serving as a proof of concept for the power and as test bench for the limitations of many currently popular theories and approaches. As such, it will not only serve as a step towards the (re-)creation of intelligence with computational means, but potentially will on a meta-level also allow to assess the suitability of current attempts at reaching this long-standing goal.

4.4. First Steps Towards an Implementation

A basic conceptual architecture for the envisioned computational framework can be sketched as presented in Fig. 2. In accordance with the program structure presented in the previous section, from a neural-symbolic perspective it consists of three main functional components with the lifting of knowledge from the subsymbolic to the symbolic level as centerpiece mediating between low-level embodied sensing and anchoring and high-level concept formation and processing.

Interaction happens both between layers within individual modules (as, e.g., between the cognitive foundations and the anchoring) as well as across components (as, e.g., through the feedback from the concept formation/reformation to the anchoring). This results in a tightly interconnected architecture forming

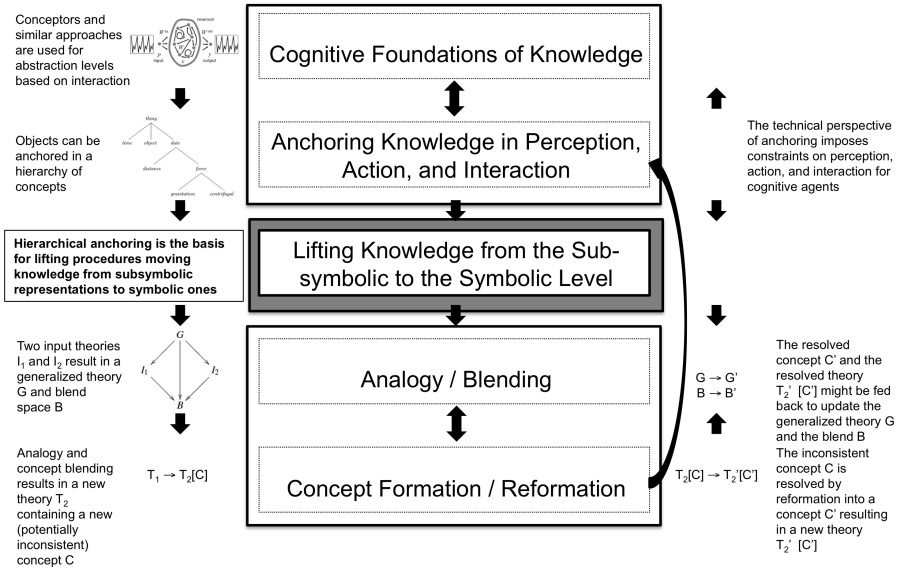

Figure 2: A schematic overview of structure and internal knowledge dynamics of the envisioned architecture featuring a closed neural-symbolic cycle of learning and reasoning (adapted from Besold et al. (2015)). While all three modules are relevant for closing the subsymbolic/symbolic cycle, the central functional component of knowledge lifting corresponds to the classical core part of neural-symbolic integration. In addition to the depicted knowledge dynamics from the level of an agent's embodied sensing and acting to conceptual theories (and back), interactions between the modules and sub-components also happen on other levels: For instance analogy-making also shall operate on the level of mechanisms during learning and knowledge lifting (see Sect. 4.3).

a dynamic whole, with changes in one layer propagating to others in order to re-establish the "harmony" introduced in Sect. 4.1.

Within the low-level subsymbolic module, the aforementioned conceptors (Jaeger (2014)), deep neural networks (Lee et al. (2009)), and similar approaches are employed in order to initially pre-structure the perceptual input stream on a subsymbolic level, augmenting the proto-structure resulting from the properties and modalities of the embodied setting. This structure can then serve as foothold for the anchoring process in a perception-based coupling of structural to environmental elements and/or to action-based percepts of the agent (also taking into account the property and attribute levels of objects/entities). Thus, while staying within the subsymbolic realm, more abstract correspondences between structured parts of the perceptual input stream and the corresponding represented content are established. These vehicle-content pairs then can be arranged in a hierarchical structure on the level of different objects/entities, but also more fine-grainedly on the level of object/entity-specific properties.

Within the high-level symbolic module, analogy and analogy-based blending are used to structure the permanently changing overall knowledge base of the agent, to transfer and adapt knowledge between similar contexts, and to create new high-level concepts through the blending of concepts in the knowledge base. These processes potentially reveal existing or introduce new inconsistencies between concepts, which can then be addressed by the top-level concept formation and reformation layer. In this highest layer, inconsistencies are fixed through manipulations of the symbolic representational structure, in turn possibly introducing new representations or concepts by altering the represented knowledge elements or the overall representation language. In order to maintain the internal structure of the overall framework, the top layer therefore might have to feed back changes or additions to the subsymbolic anchoring layer which then is forced to perform the corresponding adaptations in its assignment of objects/entities to representations.

Finally, the neural-symbolic core module in the center of the architecture bridging from low-level to high-level representations and processing builds upon

the output of the anchoring layer, i.e., correspondences between objects/entities in the perceived environment and structural elements of the subsymbolic representation, and uses deep learning techniques for representation learning in order to convert (and, by doing so, lift) the subsymbolic content-loaden representations to logic-based expressions. The corresponding learning process taps into pre-existing knowledge on the symbolic side based on the analogy mechanism and the assumption that relative temporal continuity of the environment (and, thus, the agent's input stream) should result in newly lifted symbolic concepts sharing analogical commonalities with already existing ones, which then in turn can be exploited to support the lifting process. Additionally, using the same assumption of only gradual change in the environment successive or parallel lifting processes can be implemented in a cross-informing manner, establishing analogical similarities not only over knowledge items (i.e., procedural objects) but also over the processes themselves and exploiting the expected appearance of similar sub-parts of the corresponding mechanisms.

5. Integrating Both Programs: Why the Whole Is More than the Sum of the Parts

While being at first sight almost orthogonal in approach and nature of questions (formal and theoretical on the one side, engineering-focused and systems-oriented on the other), both research programs share deep connections, have to be regarded as complementary and cross-informing, and promise to mutually augment each other in results and impact not only, but also with respect to the 'four pillars of BICA' (Stocco et al. (2010)).

The program on complexity differences and connections between different subsymbolic and symbolic formalisms for computation and representation can shed light on aspects of crucial importance for a systems-oriented program as the 'anchoring knowledge in interaction' cognitive architecture. Results from the former program can help in selecting suitable approaches within the different layers and modules of the envisioned cognitive framework assuring the feasibility

of operating the resulting implemented system. The more engineering-focused perspective of the second proposed program on the other hand provides incentive to pursue a more constructive approach to the questions asked in the theoretical research program, instead of limiting the focus to existential results. Also, the cognitive architecture offers a natural use case and empirical testbed for the expected outcomes of the complexity-oriented research endeavor, by this completing the cross-informing feedback loop between both programs.

With respect to the mentioned core lines of research on Biologically-Inspired Cognitive Architectures (i.e., the bottom-up reverse engineering of the brain, the human-like aspects of artificial intelligence, the integration of data and models, and the development of a computational architecture), the aggregate of results from both programs promises to advance significantly beyond the state of the art. As already described in Sect. 1, bridging between subsymbolic/neural and symbolic/cognitive approaches to representation and computation promises to answer several long-standing questions in the relevant fields and to establish explanatory bridges between the different perspectives on the human brain and its capacities. In doing so, especially through the second program and its agent-based embodied approach, a more human-like style of interacting in and, subsequently, learning from the environment can be expected, presumably resulting in conceptualizations and knowledge structures which are closer to human processing than current computational accounts. By closing the gap between neural representations and logic-based models the mass of data collected in the neurosciences is made accessible to use in computational-level models developed by cognitive psychologists and AI researchers, allowing for the validation or refutation of existing and the development of new hypotheses about human cognition and intelligence, while in the opposite direction also allowing for the more targeted and purpose-specific collection of new data. Finally, the second program specifically aims at delivering a cognitive architecture implementing a closed subsymbolic/symbolic acting/sensing–processing–reasoning cycle, partially building on results from the first program and pushing far beyond the state of the art in current cognitive architectures in several respects.

6. The Immediate Vision: Preparing the Ground for Really Smart Systems in the 21st Century

On the long run, integrating symbolic/cognitive and subsymbolic/neural paradigms of computation and representation is expected to solve foundational questions within AI/computer science and cognitive and computational neuroscience (as discussed in Sect. 1 and 5), at the same time bringing these fields closer to each other and establishing deeper rooted connections beyond today's level of metaphorical similarities and inspirational links between models and conceptions. At the same time, as already mentioned at the beginning of Sect. 4, successful neural-symbolic integration also promises to solve crucial challenges in neighboring fields such as representational re-description and the progressive acquisition of abstract representations from raw sensory inputs in robot architectures (Guerin et al. (2013)). There, by endowing robots with the ability to explore different ways of storing and manipulating information across the entire spectrum of subsymbolic and symbolic approaches, it shall become possible to use multiple problem solving strategies from low-level systematic search to abstract reasoning (Evans (2003)). Still, significant and lasting impact also on a shorter timescale can be expected in another domain of study and technological development, namely in the area of smart systems.

Following the advent of the internet/WWW, ubiquitous computing (Poslad (2009)) and ambient intelligence systems (Aarts & Wichert (2009)) mostly performing high-level and complex reasoning based on low-level data and signals will be key to the future development of advanced intelligent applications and smart environments. Whilst accumulating large sets of data and subsequent (often statistical) reasoning can provide for current applications Cook et al. (2009), many real-world scenarios in the near future will require reliable reasoning also based on smaller samples of data, either due to the need for immediate (re)action without the time delay or effort required for obtaining additional usable data, or due to the need of dealing with rare events offering too few similar data entries as to allow for the application of standard learning- or statistics-driven

approaches. If a bridge between the subsymbolic sensor data and high-level symbolic representations can be established, then knowledge- and rule-based approaches promise to mitigate the just described problems. Pre-coded symbolic background information and rule-based semantic processing can deal with foreseeable types of rare events, and logic-based representations of occurrences which cannot be accounted for offer the possibility to search for different, but sufficiently similar data points across different knowledge sources possibly augmenting the data set available for the low-level approaches. The corresponding systems will, thus, have to make use of subsymbolic processing side by side with complex abstract reasoning mechanisms, which then will have to be used to inform subsequent low-level sensing and processing steps in an action-oriented continuous acting/sensing–processing–reasoning cycle (similar to the cognitive system envisioned as outcome of the corresponding research program from Sect. 4).

A concrete application scenario could, e.g., be envisioned in the domain of AAL. While current systems mostly have to rely on statistical approaches and continuously growing amounts of sensor data in monitoring and interpreting the users behavior in order to operate appropriately, processing the available sensor information on all levels of the neural-symbolic hierarchy in parallel promises not only incremental progress but qualitatively new functionalities. On the one hand (mostly symbolic) semantic information could be taken into account in interpreting the observed user behavior and environment also in highly uncommon situations, allowing for previously unachieved forms of interaction: Taking an example from AAL in a care context, imagine a situation in which a cognitively-impaired person misinterprets an apple-shaped candle as an actual fruit and attempts to bite into the candle. While this constitutes a very rarely occurring setup, if the system is able to assess the high-level ontological information that, although apple-shaped, a candle does not fall into the category of eatable objects or food items this enables an intervention from the system preventing the user to proceed with the intended action. Still, full integration from the level of biologically-adequate brain models and simulations to abstract

reasoning and semantic processing (and back) in the mid-term should make even more advanced systems possible. Taking models and simulations of human perception processes and corresponding brain computations on the neural level (as, e.g., a version of Jirsa et al. (2010)'s Virtual Brain additionally equipped to also allow for external input) and making them accessible and interpretable to a smart system will eventually enable the latter to "see through the user's eyes", paving the way for a qualitatively new generation of user models. Equipped with a sufficiently accurate account of the user's (low-level) perception-based reading of the environment, subsequent biologically-inspired neural-level computations, and corresponding (high-level) interpretations, a smart system could predict when automatically induced ambient changes (such as, e.g., switching on a light source) will help clarify potential perception-based ambiguities or generally facilitate and enhance the user's perception and, thus, interactions.

Acknowledgements

The authors want to thank Artur d'Avila Garcez (City University London), Alessandro Saffiotti (Örebro University), Martin H. Fischer (University of Potsdam), and Alan Bundy (University of Edinburgh) for many helpful discussions and their active contribution in developing the described proposal for a research program on a framework and cognitive architecture for anchoring knowledge in interaction. Additional thanks go to Alexandra Kirsch (University of Tübingern) and Marcel van Gerven (Radboud University Nijmegen) for their contributions to developing the AAL application scenario sketched in Sect. 6. With respect to the research program on complexity differences between paradigms, a debt of gratitude is owed to Frank Jäkel (University of Osnabrück) for repeated rounds of feedback and many valuable comments during its inception.

References

Aarts, E., & Wichert, R. (2009). Ambient intelligence. In H.-J. Bullinger (Ed.), *Technology Guide* (pp. 244–249). Springer Berlin Heidelberg.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). Learning and Relearning in Boltzmann Machines. *Cognitive Science*, *9*, 147–169.

Adolphs, R. (2015). The unsolved problems of neuroscience. *Trends in Cognitive Science*, *19*, 173–175.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*, 373–389.

Arnold, L., Paugam-Moisy, H., & Sebag, M. (2010). Unsupervised Layer-Wise Model Selection in Deep Neural Networks. In *Proceedings of ECAI 2010: 19th European Conference on Artificial Intelligence* (pp. 915–920). IOS Press.

Bader, S., & Hitzler, P. (2005). Dimensions of Neural-symbolic Integration: A Structured Survey. In *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One* (pp. 167–194). College Publications.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, *2*, 1–127.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1798–1828. doi:`http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.50`.

Besold, T. R. (2011). *Computational Models of Analogy-Making: An Overview Analysis of Computational Approaches to Analogical Reasoning*. Technical Report X-2011-03 Institute of Logic, Language, and Computation (ILLC), University of Amsterdam.

Besold, T. R. (2015). Same same, but different? Exploring differences in complexity between logics and neural networks. In *Proceedings of the 10th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy'15)*. Neural-Symbolic.org.

Besold, T. R., Kühnberger, K.-U., d'Avila Garcez, A., Saffiotti, A., Fischer, M. H., & Bundy, A. (2015). Anchoring Knowledge in Interaction: Towards a harmonic subsymbolic/symbolic framework and architecture of computational cognition. In J. Bieger, B. Goertzel, & A. Potapov (Eds.), *Artificial General Intelligence - 8th International Conference, AGI 2015, Proceedings*. Springer volume 9205 of *Lecture Notes in Computer Science*.

Bundy, A. (2013). The interaction of representation and reasoning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *469*. doi:10.1098/rspa.2013.0194.

Chella, A., Frixione, M., & Gaglio, S. (2003). Anchoring symbols to conceptual spaces: the case of dynamic scenarios. *Robotics and Autonomous Systems*, *43*, 175–188.

Choquet, D., & Triller, A. (2013). The dynamic synapse. *Neuron*, *80*, 691–703.

Cook, D. J., Augusto, J. C., & Jakkula, V. R. (2009). Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and Mobile Computing*, *5*, 277 – 298.

Coradeschi, S., & Saffiotti, A. (2000). Anchoring symbols to sensor data: Preliminary report. In *Proceedings of the 17th AAAI Conference* (pp. 129–135). AAAI Press.

d'Avila Garcez, A., Besold, T. R., de Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K.-U., Lamb, L., Miikkulainen, R., & Silver, D. (2015). Neural-Symbolic Learning and Reasoning: Contributions and Challenges. In *AAAI Spring 2015 Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*. AAAI Press volume SS-15-03 of *AAAI Technical Reports*.

d'Avila Garcez, A., Broda, K. B., & Gabbay, D. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, *125*, 155 – 207.

d'Avila Garcez, A., Lamb, L., & Gabbay, D. (2008). *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer.

De Penning, H. L. H., d'Avila Garcez, A., Lamb, L. C., & Meyer, J.-J. C. (2011). A Neural-symbolic Cognitive Agent for Online Learning and Reasoning. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 1653–1658). AAAI Press.

Degremont, C., Kurzen, L., & Szymanik, J. (2014). Exploring the tractability border in epistemic tasks. *Synthese*, *191*, 371–408.

Diaconescu, R. (2008). *Institution-independent Model Theory*. (1st ed.). Birkhäuser.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.*, *10*, 1895–1923.

Dinsmore, J. (Ed.) (1992). *The Symbolic and Connectionist Paradigms: Closing the Gap*. Cognitive Science Series. Psychology Press.

Downey, R. G., & Fellows, M. R. (1999). *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer.

Evans, J. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science*, *7*, 454–459.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Fischer, M. H. (2012). A hierarchical view of grounded, embodied, and situated numerical cognition. *Cognitive Processing*, *13*, 161–164.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3 – 71.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Gentner, D., Holyoak, K., & Kokinov, B. (Eds.) (2001). *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press.

Gerstner, W., Kistler, W., Naud, R., & Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to to Networks and Models of Cognition*. Cambridge University Press.

Gierasimczuk, N., & Szymanik, J. (2011). A Note on a Generalization of the Muddy Children Puzzle. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge* TARK XIII (pp. 257–264). ACM.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. arXiv. 1410.5401v1 [cs.NE], 20 Oct 2014.

Guerin, F., Kruger, N., & Kraft, D. (2013). A Survey of the Ontogeny of Tool Use: from Sensorimotor Experience to Planning. *IEEE Transactions on Autonomous Mental Development*, *5*, 18–45.

Habenschuss, S., Jonke, Z., & Maass, W. (2013). Stochastic computations in cortical microcircuit models. *PLOS Computational Biology*, *9*.

Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, *14*, 1771–1800.

Hinton, G. E. (2012). A Practical Guide to Training Restricted Boltzmann Machines. In G. Montavon, G. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 599–619). Springer volume 7700 of *Lecture Notes in Computer Science*.

Immerman, N. (Ed.) (1999). *Descriptive Complexity*. Texts in Computer Science. Springer.

Isaac, A., Szymanik, J., & Verbrugge, R. (2014). Logic and Complexity in Cognitive Science. In *Johan van Benthem on Logic and Information Dynamics* (pp. 787–824). Springer volume 5 of *Outstanding Contributions to Logic*.

Jaeger, H. (2014). Controlling recurrent neural networks by conceptors. arXiv. 1403.3369v1 [cs.CV], 13 Mar 2014.

Jirsa, V., Sporns, O., Breakspear, M., Deco, G., & McIntosh, A. (2010). Towards the virtual brain: network modeling of the intact and damaged brain. *Archives Italiennes de Biologie*, *148*, 189–205.

Karpathy, A., & Fei-Fei, L. (2014). Deep Visual-Semantic Alignments for Generating Image Descriptions. arXiv. 1412.2306v1 [cs.CV], 7 Dec 2014.

Krajícek, J. (2005). Proof Complexity. In *4ECM Stockholm 2004*. European Mathematical Society.

LeBlanc, K., & Saffiotti, A. (2008). Cooperative anchoring in heterogeneous multi-robot systems. In *2008 IEEE International Conference on Robotics and Automation* (pp. 3308–3314).

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning* ICML '09 (pp. 609–616).

Lehmann, J., Chan, M., & Bundy, A. (2013). A Higher-Order Approach to Ontology Evolution in Physics. *Journal on Data Semantics*, *2*, 163–187. doi:10.1007/s13740-012-0016-7.

Leitgeb, H. (2005). Interpreted dynamical systems and qualitative laws: From neural networks to evolutionary systems. *Synthese*, *146*, 189–202.

Monner, D., & Reggia, J. (2012). A generalized LSTM-like training algorithm for second-order recurrent neural networks. *Neural Networks*, *25*, 70–83.

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*, 135 – 150.

Poslad, S. (2009). *Ubiquitous Computing: Smart Devices, Environments and Interactions*. Wiley.

Rabin, M. (1965). A simple method for undecidability proofs and some applications. In Y. Bar-Hillel (Ed.), *Logic Methodology and Philosophy of Science II* (pp. 58–68). North-Holland.

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*, 939–984.

Samsonovich, A. V. (2012). On a roadmap for the BICA challenge. *Biologically Inspired Cognitive Architectures*, *1*, 100–107.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Siegelmann, H. (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhäuser.

Sima, J., & Orponen, P. (2003). General-Purpose Computation with Neural Networks: A Survey of Complexity Theoretic Results. *Neural Computation*, *15*, 2727–2778.

Stocco, A., Lebiere, C., & Samsonovich, A. V. (2010). The B-I-C-A of Biologically Inspired Cognitive Architectures. *International Journal of Machine Consciousness*, *2*, 171–192.

Tabor, W. (2009). A dynamical systems perspective on the relationship between symbolic and non-symbolic computation. *Cognitive Neurodynamics*, *3*, 415–427.

Thomas, M., & Vollmer, H. (2010). Complexity of Non-Monotonic Logics. *Bulletin of the EATCS*, *102*, 53–82.

Uchizawa, K., Douglas, R., & Maass, W. (2006). On the computational power of threshold circuits with sparse activity. *Neural Computation*, *18*, 2994–3008.

Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books.

van Emde Boas, P. (1990). Machine Models and Simulations. In *Handbook of Theoretical Computer Science A*. Elsevier.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and Tell: A Neural Image Caption Generator. arXiv. 1411.4555v1 [cs.CV], 17 Nov 2014.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory Networks. arXiv. 1410.3916v6 [cs.AI], 7 Feb 2015.

Zhou, Z., Jiang, Y., & Chen, S. (2003). Extracting symbolic rules from trained neural network ensembles. *AI Communications*, *16*, 3–15.